A novel spectral filtering method based on optimal peak-wise response curves

Elena Novaretti
Music composer, DSP coder
elena@elenadesign.eu - Rapallo (GE) ITALY

[KEYWORDS: audio spectral filtering, peak distortion, STFT]

May, 2021

ABSTRACT

In this paper a novel artifact-free spectral filtering method suitable for realtime Short-Time Fourier Transform processing chains working with Hann-windowed frames is presented, which is not based on overlap/add, overlap/save or similar schemes and which does not introduce additional latency or phase shifts. A spectral frame of double size is composed by stacking the current and two previous overlapping frames; spectral peak bodies are identified in the resulting frame and both magnitude and phase values from the desired Impulse Response spectral curve are kept constant thru peak bodies as the filtering curve value at their local maxima; a new "peak-quantized" filtering curve of double size is thus obtained, which can in theory be multiplied directly by the double-sized frame without producing any substantial distortion to the peaks shape or to the Hann window kernel convolved around them. The resulting curve is then halved in size by linear averaging and multiplied by the Hann-windowed input frames to produce the final filtered result.

1. Introduction

Filtering is a fundamental and recurrent operation in spectral audio processing. In particular within the context of realtime Short-Time Fourier Transform (STFT) processing like the phase vocoder[1] and similar frameworks, spectral filtering is called into play either directly or indirectly everytime a modification of the spectral magnitude (or the phase, or both) of the input frames in a non-constant manner thru the spectrum is required. As it turns out, this is not just the case of mere equalization; we are facing spectral filtering when whitening a spectrum, when applying an envelope to a spectrum, when performing spectral vocoding, when subtracting a noise profile for noise removal, and whenever the very spectra have to be multiplied by other non flat spectra.

It is however a well known fact that careless algebraic multiplication of a spectrum S(w) by any function F(w) which is not a constant may lead to more or less severe artifacts.

To better understand these, it is important to visualize what happens both in Time-Domain (TD) and in Frequency-Domain (FD) when we perform a spectral multiplication.

Multiplication in FD corresponds to a circular frame-wise convolution in TD, and vice-versa. Therefore multiplying S(w) by F(w) corresponds to a circular convolution of the TD frame s(x) represented by the spectrum S(w) by the frame f(x) represented by F(w), where f(x) in our case is an Impulse Response (IR). It is evident that, unless the trivial case of F(w) = c where a flat spectrum corresponds to a pulse of one sample length at position zero, portions of the convolved signal will slide off the right end to enter back from the left end of the frame, and vice versa ("time aliasing"). In an ideal system instead, convolution should be perfectly linear: with a zero-phase IR, "future" and "past" should propagate thru convolution to the frame in exam, exactly as the content of the frame in exam ("present") should propagate to the previous consecutive frame (the "past") and to the next (the "future").

The same problem can be visualized in FD as *peak distortion*. Because of spectral leakage, a spectral peak occupies a single bin of a discrete spectrum only in the unrealistic case of an exactly bin-centered frequency; in the real world, a peak is a well defined mathematical entity - an aliased sinc function - spacing thru the whole discrete spectrum. By multiplying a peak function by a non constant value such as a frequency or phase response curve F(w), we are pretty altering its mathematical structure more or less severely depending on how abrupt the curve is, and how much it is so closer to the peak tip where its signifying information is concentrated.

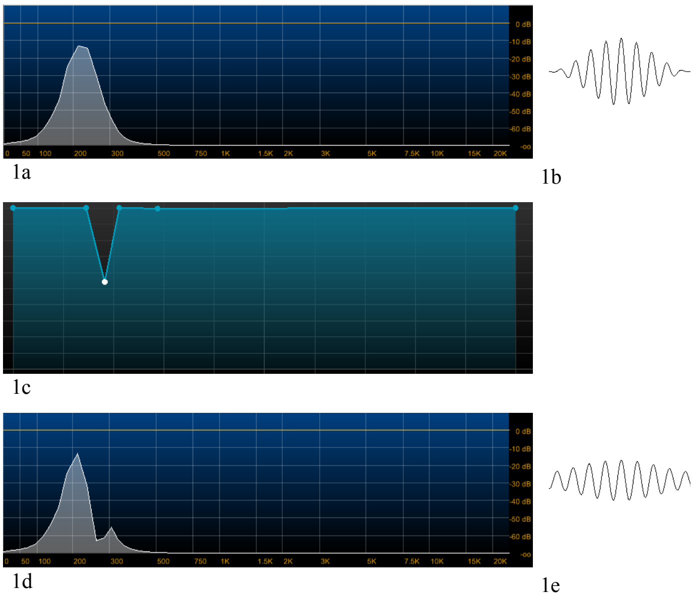
It must also be considered that the Window function used by the STFT framework, if any, like any amplitude modulation information present in TD, is stored in FD "holographically" as the spectrum of the Window function convolved around every bin. Therefore, distorcing the peak shapes will unavoidably also distort the Window function, when one is used, with the fatal consequence that the overlapping frames won't sum up to a constant audio level anymore (C.O.L.A) once added together to produce the final audio signal (see Fig.1).

To overcome the problem, high quality spectral filtering apparatuses usually resort to well established techniques known as "overlap-add" or "overlap-save"[2], aimed at converting the problem of linear convolution to a matter of circular convolution, by performing spectral multiplication on a larger frame either zero-padded or inclusive of a portion of "past" and a portion of "future".

In particular, when working with windowed frames only the second solution can realistically be adopted. Using zero-padded frames requires the window structure to be temporarily removed, which will reveal possible previous cumulating errors at the frame edges which tend to be naturally hidden otherwise, and which will be propagated inside the frame by the convolution process.

These techniques tend to be quite expensive in terms of computation power, by requiring additional both direct and inverse Fourier Transforms (FT) for every processed frame, since no methods exist to combine, split or zero-pad TD frames in FD.

But the major drawback is that by using zero-phase IRs further latency is introduced in addition to that of one frame length which is intrinsic of the STFT process. The reason is simply explained: a zero-phase IR is composed of two identical, symmetrical wings, one causal and one anti-causal. Any “future” TD sample located within the length of one IR wing past the current frame shall propagate its information back to the current frame by the convolution process. Since we cannot know the “future”, we shall forcibly introduce some latency to compensate. This is equivalent to shifting the IR to the right by the length of one wing to make it completely causal - but delayed.
On the other hand latency could be avoided converting the IR to minimum phase (MP), but even this solution comes for a price: 1. even more computation power required for the MP conversion; 2. introduction of unwanted phase shifts, which could represent a problem whenever the filtered frames were to be combined with unfiltered ones; 3. no phase filtering would be possible as a consequence.



1a

1b

1c

1d

1e

Fig. 1 *Spectral peak distortion by multiplication; a: the original spectrum; b: the original TD frame; c: the filtering curve applied; d: the resulting spectrum; e: the resulting TD frame*

2. Principle overview

We may consider a spectrum *S(w)* as the sum of many peak functions P_n of magnitude m_n i.e

$$S(w) = \Sigma\, m_nP_n(w)$$

An *ideal* spectral filter *F(w)* is expected to produce a new spectrum *S'(w)* as though every peak function in *S(w)* were re-synthesized with a new level $m'_n = m_nF(w_n)$, i.e

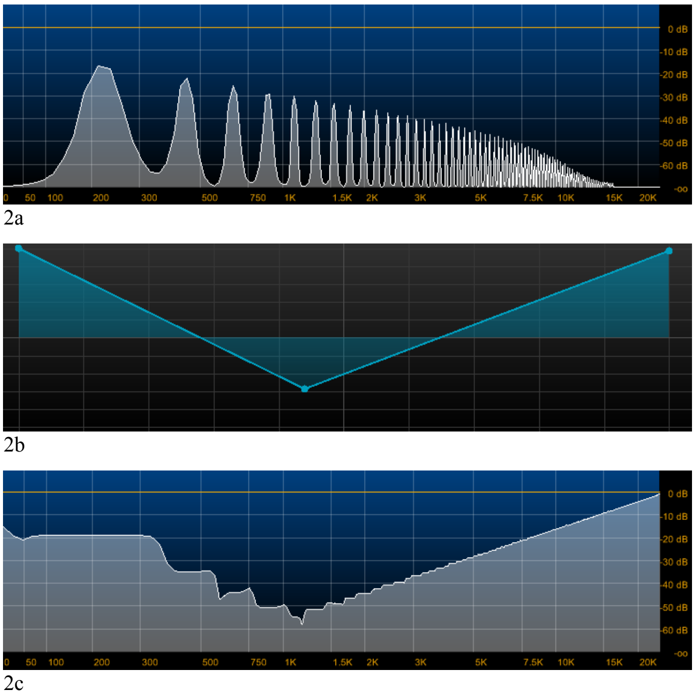
$$S'(w) = \Sigma\, m_nF(w_n)P_n(w)$$

where w_n is the bin position of the local maximum (“tip”) of P_n.
We cannot, of course, separate the various P_n in *S(w)* since they are summed together. Also, in the real world they aren't meant to be plain aliased sinc functions since they may contain amplitude modulation information convolved as explained previously, and/or they may not be perfectly stationary.

We can however synthesize a test spectrum *Z(w)* from an arbitrary set of plain peak functions of given magnitude, frequency and phase, then synthesize a second spectrum *Z'(w)* by scaling the same set of peaks by a given filter *F(w)* at w_n , and divide *Z'(w)* by *Z(w)* to obtain a new filter

$$F'(w) = Z'(w) / Z(w)$$

This filter is of capital importance because it contains some precious information. It can be deemed the *optimal* filter which, if multiplied by *Z(w)* produces exactly *Z'(w) without any peak distortion*.
By examining the magnitude curve of *F'(w)* we can easily appreciate its stepped shape, as though its value were “quantized” in a peak-wise fashion (see Fig.2). After all this is expectable, because only such a magnitude shape is able to cause no peak distortion when multiplied by the corresponding spectrum *Z(w)*, by having its local values more or less constant thru every peak body. We will refer to such a curve as “peak-wise” (PW) curve from now on.



2a

2b

2c

Fig. 2 *The optimal peak-wise filter curve; a: the original spectrum; b: the filter magnitude curve; c: the corresponding peak-wise curve*

Another method to appreciate a PW curve is processing a Hann-windowed spectral frame with an established distortion-free filtering method like the “overlap-save” scheme mentioned previously, and divide it by the original frame (taking latency into account). Once more, multiplying the original frame by the PW filter so obtained will produce the *same* distortion-free result of the overlap-save process.
This may sound as a mathematically obvious fact, but the underlying concept is fundamental: a complex and expensive process like the overlap-save can still be reduced to a plain multiplication, if we just could compute the optimal PW filter for every frame to process.

Logics suggest that we cannot compute such PW filter without knowing the future of course; but once we have understood its peak-quantized nature and how it should be qualitatively structured, we can attempt to approximate it at least, confident that the resulting errors will build up at the frame edges and therefore hidden by the window function as it is usually the case with spectral processing.

3. Algorithm in detail

Let's assume a typical realtime STFT framework generating a stream of overlapping Hann-windowed spectral frames $S_t(w)$ of nominal size z and overlap factor v; we will assume v=2 (50% overlap) in this example but any overlap factor can be used if really needed.

Parallelely, a second stream of IR frames $I_t(w)$ to be applied as filters is received, with same size and overlap factor and at the same frame clock, no matter whether dynamically changing or stationary.

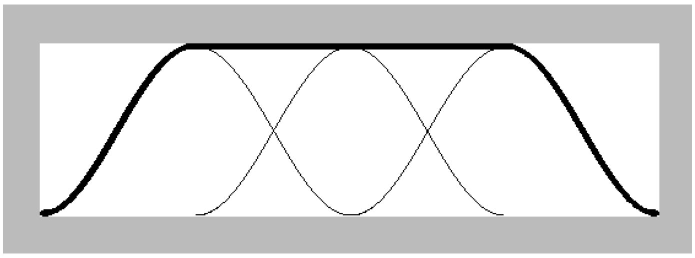


Fig. 3 *The magnitude envelope resulting by stacking three Hann-windowed frames*

Since the following operations cannot be performed efficiently in FD, the incoming frame $S_t(w)$ is first converted to TD by IFFT as $s_t(x)$. A double sized ($2z$) frame $d_t(x)$ is created by stacking (adding) the three consecutive, 50% overlapping frames s_{t-v} , $s_{t-v/2}$ and s_t each one spaced $z/2$ samples apart in TD. The resulting magnitude envelope shape $m(x)$ of $d_t(x)$ will be the one illustrated in Fig. 3, corresponding to

$$m(x)= \begin{cases} 3z/2>x>z/2 : 1 \\ otherwise : (1-\cos(2\pi x/z))/2 \end{cases}$$

always remembering that the Hann window function for a z-sized frame is

$$(1-\cos(2\pi x/z))/2$$

In order to work with a regular Hann-windowed frame with clean peaks devoid of ringing sidelobes, $d_t(x)$ is multiplied by a suitable correction curve

$$(1-\cos(\pi x/z)) \;/\; 2m(x)$$

$d_t(x)$ is then converted to $D_t(w)$ by FFT to perform peak detection: every peak body thru D_t is identified as the range [w_{min} ,w_{max}] around a local maximum w_n confined between two local minima at w_{min} ,w_{max} .

The corresponding IR frame $I_t(w)$ is doubled by linear interpolation, by simply copying every bin to a doubled bin position in the destination double IR frame $I'_t(w)$, then filling the odd bins remained empty with the average of the two neighbouring bins. (This scheme assumes dynamic filtering, i.e. a parallel stream of identically sized and overlapping IR frames; however nothing prevents the adoption of IR frames natively double in size, so that doubling and linear interpolation can be avoided.)

$I'_t(w)$ is then quantized to look like $F'(w)$ as described in section 2, by simply forcing its magnitude value (and phase if present) at every w_n to be constant through every [w_{min} ,w_{max}] interval (see Fig. 4).

A common problem with spectral analysis is inherent to overlapping peaks, i.e whenever the spectral resolution adopted cannot adequately resolve individual frequency peaks very close to each other, which tend to merge in a single peak or cluster. Because of the varying frame offset with respect to the underlying waveform, some peaks may merge in some frames but still look separated in other frames; this fact would cause the peak detection scheme just described and the resulting PW curve to return inconsistent results from frame to frame, leading to abrupt changes in the filtering magnitude and audible artifacts.

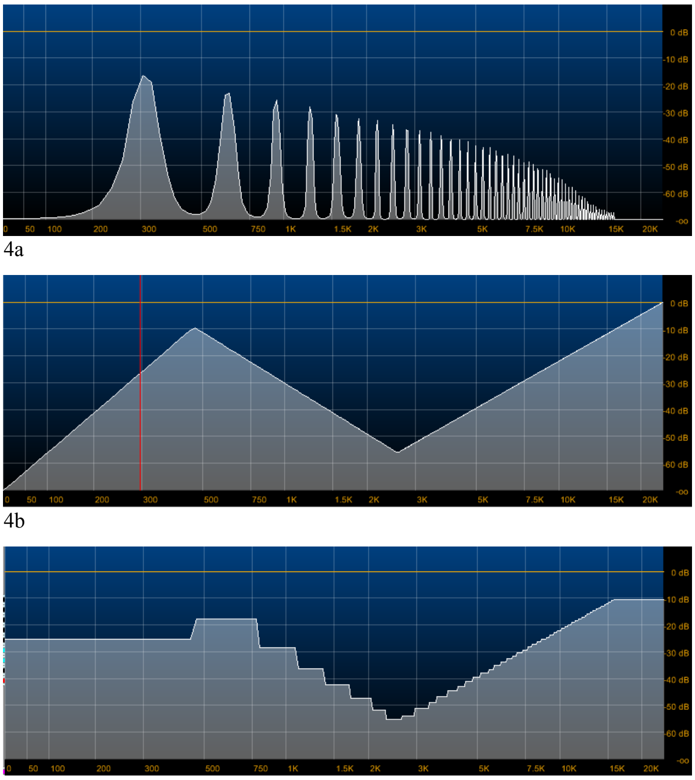
That is the fundamental reason why a double sized frame is adopted.

At this point we could simply multiply D_t by I'_t , remove the Hann window to flatten the TD magnitude envelope, keep the right half of D_t corresponding to the current frame (discarding the first half which only served the purpose of having a double frame for a better peak detection) and apply a Hann window to the result. This method proved effective indeed, but to do so we need to perform two additional FFT: one to convert $D_t(w)$ back to $d_t(x)$ in order to isolate its right half, another to convert such half to FD to provide the resulting spectral frame for futher processing. Despite optimized FFT can be very efficient on nowaday's machines, often contributing to onyl a tiny fraction of the total CPU load (being just O(N log N)) of a whole spectral processing toolchain, it is always advisable to spare resources and keep computation demands as low as possible for efficiency's sake.

Therefore an equally effective method will be adopted to provide a resulting z-sized spectral frame and which does not require any additional FFT.

I'_t is halved with average, by simply keeping all its even bins after adding the half sum of the two neighbouring odd bins (DC and Nyquist bins, by having only one neighbouring bin, will be treated differently: in their case the value of their unique neighbouring bin will be added without halving).

The resulting z-sized PW IR spectrum $I''_t(w)$ is simply multiplied by the input frame $S_t(w)$ to produce a resulting spectral frame with intact Window function and devoid of any peak distortion artifacts.



4a

4b

4c

Fig. 4 *Approximation of an ideal peak-wise filter curve (c) from a starting curve (b) relative to an input spectrum (a)*

4. Evaluation

The present work was basically motivated by the need to provide the phase-vocoder-alike modular spectral framework *Elena Design's Spectral Modules*[3] for the platform *SynthEdit*[4], developed by the Author of this work, with a reliable and robust IR-convolver module.
The present algorithm was coded in highly optimized C++ and compiled using GCC 10.2 in form of a plugin for the aforementioned framework, and tested within the intended platform on a digital audio workstation based on Intel i7 8700K @ 3.7 Ghz.
For the tests an overlap factor of 50% was used together with a fixed frame size of 2048.

Measured by the builtin SynthEdit CPU meter, peak core load was oscillating between 3 and 4% versus 6 to 7% of an equivalent reference module implementing an Overlap-Save scheme used for comparison.
Different audio material was processed by the filter even using extremely abrupt and fractured response curves, from stationary waveforms of even very low frequencies to human voice to music material. No artifacts could be detected by ear, by waveform inspection or by spectral analysis either, and the result was qualitatively identical to that offered by Overlap-Save reference module.

5. Conclusion

A novel spectral filtering method has just been exposed, which radically prevents peak distortion and disruption of the window function, based on plain algebraic multiplication by an optimal peak-wise filter curve computed as a function of a given IR filter (static or dynamic) and of every input spectral frame to be processed. The method is especially suitable and was conceived for operation within a realtime STFT framework like the phase vocoder or alike, but its core concept can easily be adapted to different scenarios. When using zero-phase IRs, no phase shift or additional latency is introduced in the process. A suitable method to compute said optimal peak-wise curve has been exposed aswell, requiring two additional FFT; it is however still matter of study whether simpler means to do the same exist, requiring even less computations or no FFT at all. It is also matter of investigation whether an even better estimate of the "true" peak-wise filter *F'(w)* as explained in section 2, and not just an approximation, is possible at all and with simple means.
The novelty of the present method however is inherent to its purely frequency-domain nature, where time-domain aliasing and alteration of the Window structure are prevented by maintaining the frequency-domain filter curve value constant through the whole extension of every spectral peak body. In time-domain the whole process can be seen as the convolution by an IR which dynamically adapts to the frame content to substantially prevent the aforementioned distortions. A spectrum *close* to that offered by the "ideal" filter described in section 2 is produced, i.e as though it were synthesized with peak functions of the desired magnitude, and without all the time-domain implications inherent to spectral filtering.
However the whole procedure must still be deemed an approximation, because only a complete processing scheme like the Overlap-Save or Overlap-Add can offer a mathematically "perfect" result, despite with their many drawbacks. This limitation arises from the peak detection method adopted and the respective quantization of the filtering curve, where peak structures are considered only partially, i.e within their non-overlapping (obvious) portions. The resulting errors, however, as it is the case with most approximate spectral operations, tend to build up at the frame edges; it is therefore imperative the usage of Hann windowing and overlapping frames to make such errors pretty disappear, by working with frames containing their signifying information concentrated around the center.
In real world tests the algorithm performed excellently with both stationary audio signals of even very low frequency and with music material of various kind, with a quality identical to that offered by the standard Overlap-Save method and no detectable artifacts.

References

[1]. Flanagan J.L and Golden, R.M. (1966) "Phase Vocoder". Bell System Technical Journal 45 (9): 1493-1509

[2]. Rabiner, Lawrence R.; Gold, Bernard (1975), "Theory and application of digital signal processing", pp. 63-67

[3]. Elena Spectral Modules for SynthEdit
https://www.kvraudio.com/product/elena-spectral-modules-for-synthedit-by-elena-design

[4]. SynthEdit, visual modular developing environment for virtual sound synthesizers and effects - www.synthedit.com